Quantifying the Uncertainty of Age-Specific Mortality Estimates in Data-Scarce Contexts

Nathaniel Darling, Yuan Cui, Irena Chen, Ugofilippo Basellini, Monica Alexander

September 2024

Abstract

In many countries, the lack of vital registration systems means that age-specific mortality rates must be derived from statistical models or model life tables. In such settings, child mortality is often available from targeted surveys. State-of-the-art methods draw on the fundamental relationship between child mortality and mortality at other ages to predict the latter when only the former is known. These methods currently provide point estimates for the entire age-pattern of mortality, but with no quantification of uncertainty. The contribution of this paper is to distinguish four sources of uncertainty which affect estimates of mortality schedules, and to extend two state-of-the-art models to account for such uncertainty. We extend and fit these models in a Bayesian framework, and derive estimates of age-specific mortality with credible intervals. We expect to find that the estimated credible intervals depend on the available data quality, the age group being estimated, and the statistical assumptions.

1 Introduction

In many countries, concentrated in low- and middle-income regions around the world, vital registration systems are non-existent or incomplete, so most or all deaths go unrecorded (Karlinsky, 2024). In these countries, age-specific mortality rates for all age groups are generally derived from model life tables or statistical models. In such settings, information about child mortality is often available from targeted surveys such as the Demographic and Health Surveys. State-of-the-art methods thus draw on the fundamental relationship between child mortality and mortality at other ages to predict the latter when only the former is known (Wilmoth et al., 2012; Clark, 2019).

Two common methods for predicting the entire age-specific mortality schedule from child mortality are the Log-Quad and SVD-comp models. The Log-Quad model (LQ) predicts mortality rates at each age group as a quadratic function of the log of child mortality (${}_5q_0$) (Wilmoth et al., 2012). In this approach, a separate model with three parameters is fitted to each age group. The LQ approach is parsimonious and performs well under a range of settings, and has since been extended to predict patterns of child mortality early ages (Guillot et al., 2022). The SVD-comp method (SVD) developed by Clark (2019) decomposes a set of life tables into four principal components, and estimates the loadings on each of these components as a function of child mortality. Both SVD and LQ approaches can optionally incorporate adult mortality, if available.

Each of these approaches has been shown to outperform traditional model life tables in terms of predictive error (Wilmoth et al., 2012; Clark, 2019). These approaches also have the advantage of being transparent and relatively simple to implement. A drawback with both approaches is that they currently provide point estimates for age specific mortality rates for each country year, with no quantification of uncertainty.

In this project, we aim to extend existing methods of estimating age-specific mortality to account for uncertainty in the resulting estimates. Quantifying uncertainty is preferable over a simple point estimate and is particularly important in data-scarce contexts. Most fundamentally, a point estimate constitutes a loss of information compared to a full uncertainty interval. Uncertainty intervals show the range of possible predictions that are consistent with the data and the model, and the relative size of the uncertainty intervals reflect the quality and reliability of estimates in particular contexts. The users of the estimates can then assess whether this uncertainty is large or small in the context that the estimates are being used for. Further, uncertainty intervals make clear that what is being presented is a modelled estimate, rather than ‘data’ or ‘truth’. While demographic estimates have traditionally been presented without estimates of uncertainty, the past decade has seen a substantial shift in producing and reporting probabilistic estimates, driven partly by the move of the United Nations Population Division to produce probabilistic population projections as part of the World Population Prospects (Raftery et al., 2012)

We distinguish four sources of uncertainty that should be considered.

1. Uncertainty due to measurement error in predictor variables (child mortality). As the child mortality estimates are derived from surveys, these inputs have associated sampling error, as well as potential other sources of error, such as recall bias (Alkema and New, 2014). Failing to account for measurement error can potentially bias predictions and reduce statistical power to detect an effect (Carroll et al., 2006).

2. Uncertainty that comes from missing data. In particular, both the LQ and SVD models can be fit in settings where both child and adult mortality are available, or in settings where only child mortality is available. Uncertainty should be greater in settings where fewer predictors are available.

3. Uncertainty in the coefficient estimates of the model. The LQ and SVD assume a log linear or log quadratic relationship, and the coefficient estimates dictating this relationship may have differing amounts of uncertainty, based on the context of interest.

4. Uncertainty from stochastic elements of model: for example, assuming that the underlying mortality process itself is a stochastic process rather than a fixed quantity. This type of uncertainty has a larger impact on estimates for smaller populations, where stochastic variability is higher.

Our approach to uncertainty quantification will incorporate all the first four sources of uncertainty. We will reformulate models in a Bayesian setting, which provides a natural framework for incorporating different types of uncertainty and propagating this uncertainty to the final estimates. In particular, in Bayesian inference the posterior predictive distribution incorporates sources 3 and 4, and the model can be naturally extended to explicitly model uncertainty stemming from sources 1 and 2.

The contribution of this paper is to augment the LQ and SVD models by providing credible intervals for estimates of mortality. We fit the models in a Bayesian framework, which further allows us to estimate the LQ model simultaneously for all age groups. We validate the models not only by comparing point estimates to observed values, but also by comparing the credible intervals to the data. Future work will also focus on extending existing models to improve age-specific mortality estimation.

2 Data and methods

The data used in this paper are drawn from the Human Mortality Database (2024). We use all available data, which covers 41 countries over the period 1751-2023, corresponding to a total of 4906 unique country-year observations.[1] Throughout the paper, 1x1 (age group x year) data are used.

We develop and compare two models: a version of the Log-Quad, and the SVD-comp. The models are described below.

2.1 Log-Quad model

First, we adapt the Log-Quad Model for single years of age. In particular, we assume that the probability of death at age x for a particular country-year i, q_{xi} is log-normally distributed with a mean and variance:

$$\log(q_{xi}) \mid \mu_{xi}, \sigma_{xi} \sim N(\mu_{xi}, \sigma_x^2) \tag{1}$$

[1] Most countries do not have observations for the entire period. For some countries, life-tables are reported for sub-populations, so the 41 countries correspond to 50 sub-populations.

The mean parameter for each age and country-year μ_{xi} has the form:

$$\mu_{xi} = a_x + b_x \log({}_5q_{0i}) + c_x \log({}_5q_{0i})^2 \tag{2}$$

where ${}_5q_{0i}$ is the observed under-five child mortality for country-year i and a_x, b_x, and c_x are parameters to be estimated.

We model the stochastic element of the model as normally distributed with a different variance σ_x for each age group x, which we estimate from the data. Allowing the variance to variance by age accounts for the the fact that the log-quadratic relationship between child mortality and mortality at age x is more variable for some ages. Note that currently, we are not accounting for measurement error in the data inputs ${}_5q_{0i}$. Future work will focus on extending the model to account for varying amounts of error in the covariates.

We estimate models for each age group simultaneously, and pool information across age groups by relating the parameters a_x, b_x, and c_x through a random walk structure (Rue and Held, 2005). Specifically, we use a non-informative prior for $x = 0$: a_0, b_0, $c_0 \sim N(0, 100)$. For all other age groups, $x \neq 0$, we use the following model:

$$a_x \sim N(a_{x-1}, \sigma_a^2) \tag{3}$$

$$b_x \sim N(b_{x-1}, \sigma_b^2) \tag{4}$$

$$c_x \sim N(c_{x-1}, \sigma_c^2) \tag{5}$$

We place a non-informative prior on each σ_x (Gelman, 2014):

$$\sigma_x \sim \text{Gamma}(1, 5) \tag{6}$$

Posterior samples of all parameters are obtained using Gibbs sampling, a Markov Chain Monte Carlo algorithm, performed using JAGS software in R (R Core Team, 2024; Plummer, 2023). Posterior samples of μ_{xi} and σ_{xi} are used to draw new samples of $\log(q_{xi})$ from posterior predictive distribution. These represent predictions of age-specific mortality in a country-year of interest. Uncertainty in these estimates is reported as a 95% credible interval, based on calculating the 2.5th and 97.5th quantiles of the posterior predictive distributions.

2.2 SVD model

Following the procedure described by Clark (2019), we arrange HMD life-tables into an 110 × 4906 matrix of (age groups by country-years in the HMD). Each element of the matrix contains the observed logit transformed probability of death in the corresponding year and life-table.

We use singular-value decomposition to represent the mortality schedule for each country as a weighted sum of the scaled left singular vectors:

$$\boldsymbol{q_i} \approx \sum_{c=1}^{C} v_{ic} \cdot s_c \boldsymbol{u}_c \tag{7}$$

where $\boldsymbol{q}$ is a mortality schedule across ages, i indexes country-years, C refers to the number of components, c indexes components, V are the right singular values, s are the scaled values, and $\boldsymbol{u}$ are the vector of left singular values. Using a low value of C (of 4) can account for the vast majority of variance in the HMD data.

Whereas s and u are constant across all mortality schedules, v varies by country year. For a country where it is not known (because the full mortality schedules are not available), it can be estimated on the basis of other covariates.

We therefore replace v_{ic} with $\tilde{w}_c$, an estimated weight:

$$\hat{\boldsymbol{q}} = \sum_{c=1}^{C} \tilde{w}_c \cdot s_c \boldsymbol{u}_c \tag{8}$$

We estimate weights assuming that they are normally distributed:

$$\tilde{w}_{ic} \mid \mu_{wic}, \sigma_{wc} \sim N(\mu_{wic}, \sigma^2_{wc}) \tag{9}$$

The mean parameter for each weight and country-year μ_{wic} has the form:[2]

$$\mu_{wic} = \alpha + \beta_1 \log({}_5q_{0i}) + \beta_2 \log({}_{45}q_{15i}) \tag{10}$$

where ${}_5q_{0i}$ is the observed under-five child probability of death for country-year i, ${}_{45}q_{15i}$ is the observed probability of death between ages 15 and 45 for country-year i, and α_1, β_1, and β_2 are parameters to be estimated.

We use a Bayesian model which estimates the loadings on each of the first four principal components from child and adult mortality. Following Clark (2019), where adult mortality is not available, it is estimated from child mortality. We assume that ${}_{45}q_{15i}$ is log-normally distributed:

$$\log({}_{45}q_{15i}) \mid \mu_p, \sigma_p \sim N(\mu_p, \sigma^2_p) \tag{11}$$

The mean parameter for each adult mortality rate has the form:[3]

$$\mu_p = \alpha_p + \beta_p \log({}_5q_{0i}) \tag{12}$$

where ${}_5q_{0i}$ is the observed under-five child probability of death for country-year i and α_p and β_p, are parameters to be estimated.

In this Bayesian setting, the additional uncertainty from estimating adult mortality when it is not observed is propagated forward into the posterior predictive distribution. As in the LQ model, we use non-informative priors of Normal distributions with wide variance for regression coefficients, and Gamma distributions for variance parameters.

3 Preliminary findings

Figure 1 shows the relationships between child and adult mortality at different ages, and visualises the relationships underpinning the LQ model. At younger ages, child mortality is more strongly predictive of mortality. However, there remains a large amount of variance in q_x (on the log scale) for any given level of adult mortality. This means that credible intervals in relative terms remain very large at younger ages, reflecting the considerable variance in mortality in the young in relative terms across historical countries and time periods, even after conditioning on the level of child mortality.

At older ages, child mortality is a weaker predictor of q_x. However, there is low variance (on the log scale) across observations of q_x at older ages, meaning that our credible intervals remain small in relative terms. (Though in absolute terms, intervals are highest at older ages, because the base rates are much higher).

Figure 2 shows mortality estimates for two countries, Norway and Chile, in 2022. It can be seen that the credible intervals in the LQ model are much larger at lower ages than at higher ages (on the log scale), reflecting the greater relative variability of mortality rates at these ages in the data. Under the SVD model, however, whilst the point estimates are very close to those derived from the LQ model, the pattern of credible intervals are different. In particular, they vary much less by age, reflecting the fact that the error term was only enabled to vary by component, rather than by age, in the SVD model. Further work will calibrate the credible intervals by exploring alternative specifications of the error terms, and by comparing the credible intervals with the observations.

[2] We are using a simpler functional form than the estimating equation used by Clark (2019). We will explore other specifications of this relationship.

[3] Again, we are using a simpler functional form than used by Clark (2019), but we will explore other specifications.

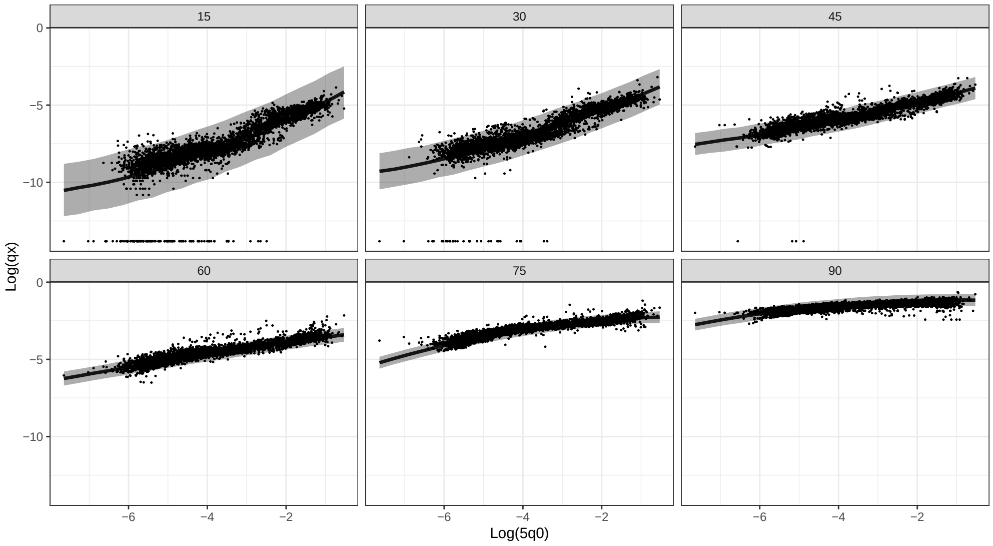

Figure 1: For selected ages, we plot the relationship between $\log({}_5q_0)$ and q_x. We show observations (points), fitted values under the LQ model (solid line) and 95% credible interval (dashed line).

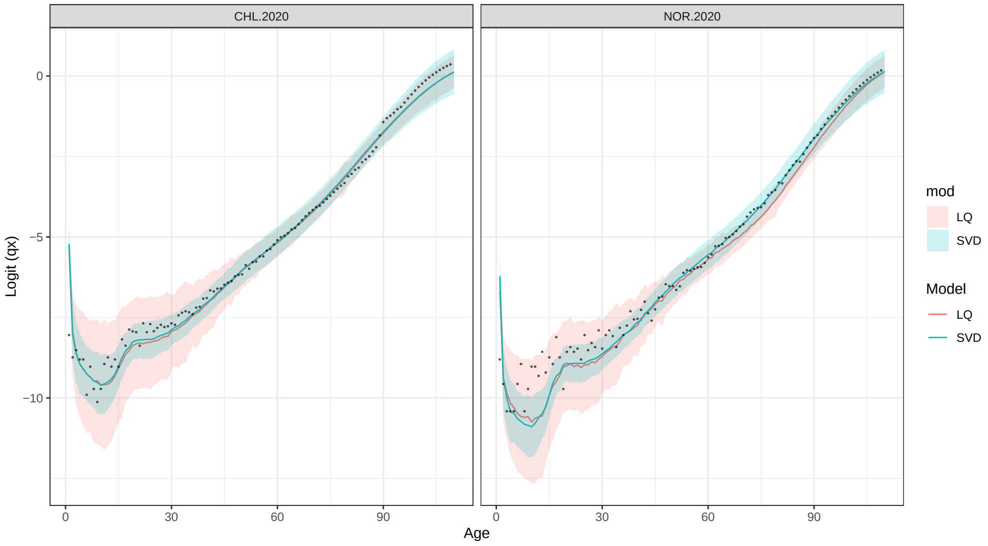

Figure 2: Point estimates (solid lines), 95% credible intervals (ribbons) and observed values (dots) for the age-specific mortality schedules for two country years, predicted based on child and adult mortality for LQ and SVD models.

4 Further directions

This preliminary work outlined here demonstrates how the flexibility of Bayesian modelling can enable the quantification of uncertainty in estimates of mortality rates. The full version of this paper will additionally incorporate the following analyses:

1. Systematic analysis of the coverage of credible intervals, as part of a process of model calibration and validation on out-of-sample test datasets.

2. Extending the comparisons of uncertainty in the LQ and SVD models in Figure 2 to alternative model specifications and across a range of values of predictor variables.

3. Explicit discussion and modelling of measurement error in child mortality, arising from survey design (Alkema and New, 2014; Alexander and Alkema, 2018).

4. Exploration of the effect on point estimates and uncertainty which comes from adult mortality rates being unavailable in some settings.

5. Relaxing the quadratic relationship of LQ model in favour of a smoother non-linear relationship, and similarly exploring different functional forms for estimating weights from covariates in the SVD model.

6. Seeking to use other information to improve model predictions, in particular, the neonatal mortality rate. There seems to be potential to leverage modeling efforts and findings from the child mortality log-quad model (Guillot et al., 2022) to extend existing models for adult mortality to be able to better discriminate between different mortality regimes.

7. The application of these models to estimating mortality rates in subnational areas in low- and middle-income countries, drawing on subnational child mortality estimates produced by the UN Inter-agency Group for Child Mortality Estimation.

8. Developing an open source user-friendly software package which will enable other researchers to easily implement these models.

References

Alexander, M. and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, 38:335–372.

Alkema, L. and New, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline Bias-reduction model. *The Annals of Applied Statistics*, 8(4):2122–2149.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* CRC Press LLC.

Clark, S. J. (2019). A general age-specific mortality model with an example indexed by child mortality or both child and adult mortality. *Demography*, 56(3):1131–1159.

Gelman, A. (2014). *Bayesian data analysis / Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin.* Texts in statistical science.

Guillot, M., Romero Prieto, J., Verhulst, A., and Gerland, P. (2022). Modeling Age Patterns of Under-5 Mortality: Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data. *Demography*, 59(1):321–347.

Human Mortality Database (2024). University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de. Data retrieved on September 27, 2024.

Karlinsky, A. (2024). International completeness of death registration. *Demographic Research*, 50(38):1151–1170.

Plummer, M. (2023). *Rjags: Bayesian Graphical Models Using MCMC.*

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* Vienna, Austria.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., and Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109(35):13915–13921.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* Chapman and Hall/CRC.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., and Sawyer, C. (2012). A flexible two-dimensional mortality model for use in indirect estimation. *Population studies*, 66:1–28.